Consistency checking in a cloud of interlinked Cultural Heritage knowledge graphs – first results of using the SampoSampo data service and portal

Petri Leskinen[1,*,†], Annastiina Ahola[1,†], Heikki Rantala[1,†], Jouni Tuominen[2,1,†] and Eero Hyvönen[1,†]

[1]*Semantic Computing Research Group (SeCo), Aalto University, Finland*
[2]*Helsinki Institute for Social Sciences and Humanities (HSSH), University of Helsinki, Finland*

Abstract

This paper presents a novel use case and functionalities of the new data linking service and portal *SampoSampo – Connecting Everything to Everything Else* on top of it, based on a Linked Open Data (LOD) cloud of related Cultural Heritage (CH) knowledge graphs (KG) of different application domains. This new Sampo system interlinks data about historical Finnish people, organizations, places, and events from 11 earlier Sampo systems in use on the Web as well as 8 other Finnish and international sources of data on the Web. This paper focuses on one particular use case of SampoSampo: how to detect automatically conflicting and complementary data about entities in an interlinked cloud of CH KGs. With provenience information attached, this is useful 1) for data publishers for checking the quality and possible errors in their data by comparing it with datasets provided by the other publishers, and 2) for data consumers for verifying the quality of data based on multiple primary sources. Our first results show some striking conflicts and disagreements between the primary data sources interlinked in SampoSampo.

Keywords

linked data, digital humanities, entity alignment, data validation, semantic portal, data analysis, knowledge discovery

1. SampoSampo – Connecting Everything to Everything Else

SampoSampo – Connecting Everything to Everything Else[1] (Hyvönen, Ahola, Leskinen, Rantala, et al. 2025; Hyvönen, Ahola, Leskinen, and Tuominen 2025) is new member in the Sampo series on Linked Open Data (LOD) services and portals (Hyvönen 2022) in use in Finland. In contrast to earlier Sampos[2], it is a "metasampo" based on a cloud of 11 other Sampos (BiographySampo, AcademySampo, LetterSampo Finland, etc.) and 8 related external data services on the Web (Wikidata, Geni.com, Getty ULAN, etc.). The SampoSampo knowledge graph is a data linking service with resemblance especially to the viaf.org[3] service (Hickey and Toves 2014) provided by OCLC, but also to the works of ontology mapping, ontology services (Xia, Jiménez-Ruiz, and Cross 2015; Frosterus et al. 2015; Laouenan et al. 2022), Linked Open Vocabularies[4], and the proxy data model of Europeana (Isaac 2023).

DHNB 2026: Lost in Abundance: Encounters with the Non-Canonical, March 9–13, 2026, Aarhus, Denmark
*Corresponding author.
†These authors contributed equally.
petri.leskinen@aalto.fi (P. Leskinen); annastiina.ahola@aalto.fi (A. Ahola); heikki.rantala@aalto.fi (H. Rantala); jouni.tuominen@helsinki.fi (J. Tuominen); eero.hyvonen@aalto.fi (E. Hyvönen)
https://www.aalto.fi/en/people/petri-leskinen (P. Leskinen); https://www.aalto.fi/en/people/annastiina-ahola (A. Ahola); https://seco.cs.aalto.fi/u/rantalh3/ (H. Rantala); https://seco.cs.aalto.fi/u/jwtuomin/ (J. Tuominen); https://seco.cs.aalto.fi/u/eahyvone (E. Hyvönen)
0000-0003-2327-6942 (P. Leskinen); 0009-0008-6369-4712 (A. Ahola); 0000-0002-4716-6564 (H. Rantala); 0000-0003-4789-5676 (J. Tuominen); 0000-0003-1695-5840 (E. Hyvönen)

Digital Humanities in the Nordic and Baltic Countries Publications – ISSN: 2704-1441

[1]SampoSampo project homepage: https://seco.cs.aalto.fi/projects/ss/
[2]Sampo series of over 20 LOD services and CH portals: https://seco.cs.aalto.fi/applications/sampo/
[3]Virtual International Authority Files: https://viaf.org
[4]Linked Open Vocabularies: https://lov.linkeddata.es/dataset/lov/

The goal of viaf.org is "to lower the cost and increase the utility of library authority files by matching and linking widely-used authority files and making that information available on the Web". In contrast, our work demonstrates how such a KG and LOD service can be used for the following practical purposes: 1) How to do semantic search and browsing on a global level over all KGs in the cloud. 2) How to detect how the data about the same entities is conflicting or complementary in different KGs of the cloud. 3) How do discover new relations with explanations between entities over the KG cloud (a form of explainable AI (Došilović, Brčić, and Hlupić 2018)). 4) How to use the linking service in creating new data services and applications, in our case new Sampos.

This paper presents the first results of using SampoSampo in the case (2) above using the SampoSampo portal demonstrator and Google Colab scripting for data analysis. As a novelty, our data service facilitates analyses about linkedness of the datasets by shared entities and finding conflicting data in the interlinked datasets. Biographical data about historical people and places are considered as a case study. The analyses presented are useful for both data publishers and end users for quality checking that is necessary for reliable historical research. Our first results show that conflicts and disagreements are fairly common even between primary data sources that are in general considered reliable.

2. Finding Conflicting Data in a Cultural Heritage KG Cloud

As customary in Sampo systems, there are two ways of analyzing the underlying KG: 1) By accessing the SPARQL endpoint directly by scripting and tools, such as Google Colab and Yasgui editor, or 2) using the Sampo portal where programming skills are not needed, only data literacy:

2.1. Data analysis using the SPARQL endpoint and Google Colab

The focus of the work presented in this article is to find conflicting information about the biographical people and geographical places in the data sources of a KG cloud. The SampoSampo dataset contains approximately 96 800 person and 28 800 place entities in total. The number of datasets that provide information about a single biographical person is depicted in Figure 2.1. For example, 32 873 people have their data coming from two distinct data sources. On average, each individual's information was sourced from 2.63 different databases. In total, 16 100 individuals or 16.6% of the entire dataset were represented in only a single database. Conversely, only the most prominent people appeared in more than 12 databases, with the highest instance being found across 16 different sources. For the place data the average number of data sources was 1.83 and 50.6% of the data had a single source of data.

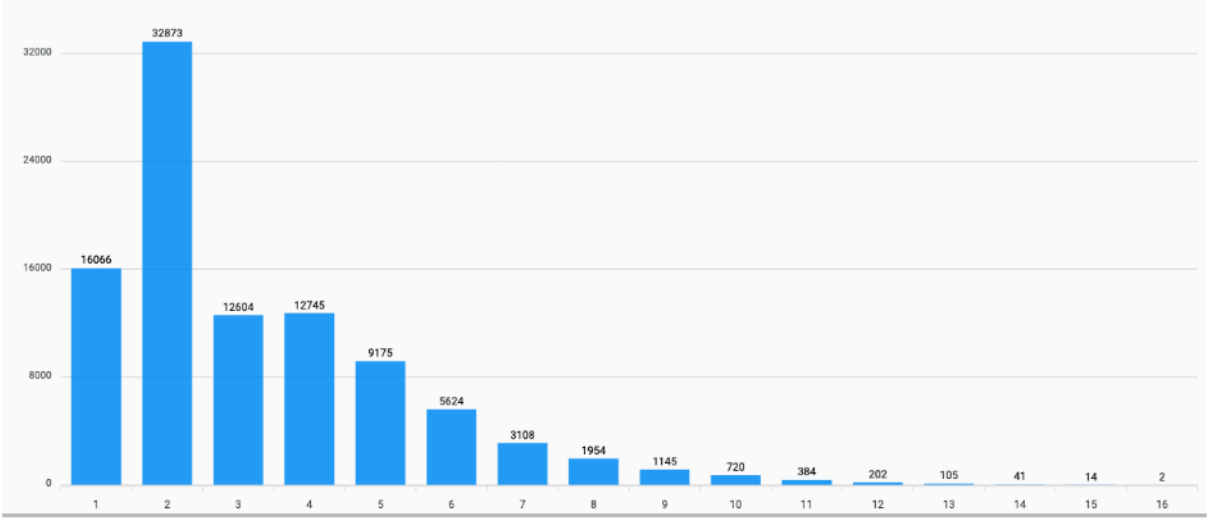

Figure 1: Distribution of the number of datasets for a single person entity in SampoSampo dataset

The analyzed properties for biographical data were the time and places of birth and death as well

as the gender of the individuals. Missing values from the data sources were not considered as errors. Instances where the times of birth or death were expressed with different levels of precision—such as a specific date in one source compared to only a year in another—were common. These variations were not considered errors but rather artifacts of differing granular standards used in data sources. Similarly, for the places the different resolutions were not considered as mistakes, e.g., the mentioned location could refer to a single mansion, a village or town, or entire municipality or county in different data sources. The few cases of inconsistent genders are mostly caused by either having a unisex given name or by using a nickname or pseudonym referring to the opposite gender when, e.g., publishing her or his literary works. Given the diversity in name spellings and usage of nicknames or pseudonyms, we chose not to evaluate variations in the given and family names.

Evaluation results of analyzing the biographical data are depicted in Table 1. Each of the evaluated properties are given three measures, the total (#Total) count of values extracted from the data source, the number of inconsistent values (#Inc.) and the percentage of inconsistent values (%) with respect to the total count. Each row corresponds to one of the 17 data sources with the total count of entire data given at the bottom row. The properties that were not provided in the data source are marked with hyphen.

Table 1
Overview of inconsistencies in biographical data

	Place of birth			Place of death			Time of birth			Time of death			Gender		
	#Total	#Inc.	%	#Total	#Inc.	%	#Total	#Inc.	%	#Total	#Inc.	%	#Total	#Inc.	%
KANTO	9588	236	2.46	9167	80	0.87	31615	359	1.14	31638	208	0.66	-	-	-
Norssi	683	13	1.90	232	7	3.02	701	35	4.99	701	39	5.56	701	0	0.00
Geneanet	31593	564	1.78	28010	677	2.42	41435	2750	6.64	41435	2140	5.17	41395	5	0.01
BiographySampo	11853	180	1.52	11564	168	1.45	21527	759	3.53	21527	386	1.79	21493	2	0.01
BookSampo	4165	60	1.44	3332	64	1.92	9154	194	2.12	9153	180	1.97	9150	10	0.11
Wikidata	23661	340	1.44	21810	200	0.92	40139	1419	3.54	40139	1053	2.62	40139	20	0.05
ParliamentSampo	2089	19	0.91	1964	36	1.83	2093	3	0.14	2093	3	0.14	2093	1	0.05
ArtSampo	776	6	0.77	681	9	1.32	1106	15	1.36	1106	8	0.72	1084	2	0.18
AcademySampo	21090	153	0.72	21893	238	1.09	22675	682	3.01	22675	574	2.53	22675	1	0.00
WarSampo	3320	13	0.39	1709	22	1.29	4156	27	0.65	4156	39	0.94	-	-	-
LetterSampo	17492	63	0.36	16717	81	0.48	31319	271	0.86	31319	257	0.82	31319	4	0.01
OperaSampo	-	-	-	-	-	-	575	15	2.61	567	7	1.24	550	1	0.18
Wikipedia	-	-	-	-	-	-	20558	526	2.56	20558	516	2.51	-	-	-
ULAN	-	-	-	-	-	-	1847	44	2.38	1847	45	2.44	1847	7	0.38
Snellman	-	-	-	-	-	-	3223	32	0.99	3223	18	0.56	-	-	-
Edelfelt	-	-	-	-	-	-	3296	16	0.48	3328	11	0.33	-	-	-
HISTO	-	-	-	-	-	-	536	1	0.19	536	0	0.00	-	-	-
All datasets	128342	1723	1.34	119029	1629	1.37	239226	7391	3.09	239272	5669	2.37	238454	54	0.02

2.2. Data analysis using the SampoSampo portal

The SampoSampo portal includes application perspectives for searching entities, including people and places, in all underlying CH KGs at the same time, or in a subset selection based on a dataset selection facet. For each entity, an instance page is generated that aggregates information about the entity from the different interlinked data sources with provenience information link attached that tells the source of the different pieces of information. In addition, a specific "Inconsistency facet" was automatically created that can be used to find and analyze various types of inconsistencies in data easily with one click. For example, in the People search perspective this facet includes the following categories with the number of hits in brackets: Time of birth [7391], Time of death [5669], Place of birth [1723], Place of death [1629], gender [54]. For example, the category "Time of birth [7391]" tells that there are some disagreements or issues of precision about the time of birth regarding 7391 people in the primary datasets.

Figure 2.2 depicts a screenshot from the portal on the instance page of the prominent Finnish bishop Mikael Agricola (1510–1557), called the "father of the Finnish language". The inconsistencies with the times of birth and death and the place of death are marked with red showing also the corresponding data sources, such as Wikidata and Booksampo, as superscript. Since finding out the ground truth would often require close reading by a domain expert, the wrong value and data source is concluded to be the one appearing least of the times, like the time of death 1557–03–30 stated in Wikidata while

Preferred labels per proxy	Kanada ... AcademySampo, BiographySampo, BookSampo, LetterSampo, ParliamentSampo, Wikidata, YSO
Alternative labels per proxy	Canada ... AcademySampo, BiographySampo, LetterSampo, ParliamentSampo, YSO
Latitude	• 56.0 LetterSampo, Wikidata, YSO • 56.130366 AcademySampo, BiographySampo • 60.10867 BookSampo, ParliamentSampo
Longitude	• -106.346771 AcademySampo, BiographySampo • -109.0 LetterSampo, Wikidata, YSO • -113.64258 BookSampo • -113.6426 ParliamentSampo
Sampled latitude	56.0
Sampled longitude	-109.0

Figure 2: Screenshot from the portal showing the different coordinates for Canada in the data.

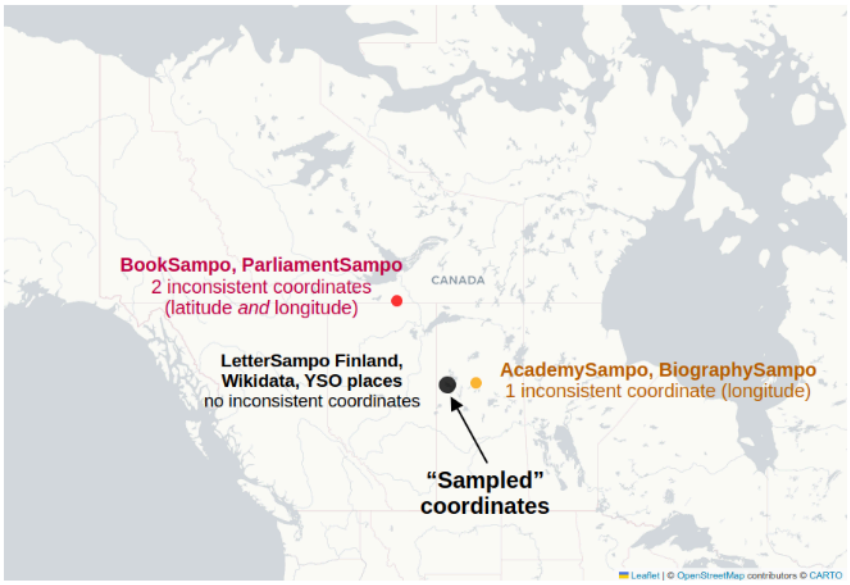

Figure 3: Canada's different coordinates in the data visualized on a map. (Leaflet, © OpenStreetMap contributors, © CARTO)

the other four datasets provide either the year 1557 or an exact date 1557–04–09. In the case of place of death, there have been several places in Finland called *Uusikirkko* (Newchurch), one of them later renamed to *Kalanti*.

Developing the person ontology of SampoSampo was carried out in the first place as a part of the LetterSampo Finland project (Hyvönen, Leskinen, et al. 2025). During the process issues like choosing the correct preferable label for an individual were tackled and the practice of showing the data variations inheriting from chosen sources of data—including the discrepancies—was adapted.

For the geographical entities the differences in latitude and longitude of given geocoordinates were analyzed. Each of the latitude and longitude values were compared against the "sampled" coordinates—latitude and longitude values from the proxies of the datasets that on average have the smallest Euclidean distances to all the other coordinates—computed for the provided place entities in SampoSampo data. If the values had a difference of more than 0.5 decimal degrees compared to their respective "sampled" values, they were marked as inconsistent.

Table 2 depicts the evaluation measures about inconsistencies for places in Wikidata and the primary data underlying some Sampo systems. It is important to note that having inconsistent coordinates for place does not always indicate an error in the given coordinates, as the data includes places of varying geographical area sizes. For example, as shown in Figure 2.2, the entity for the country of

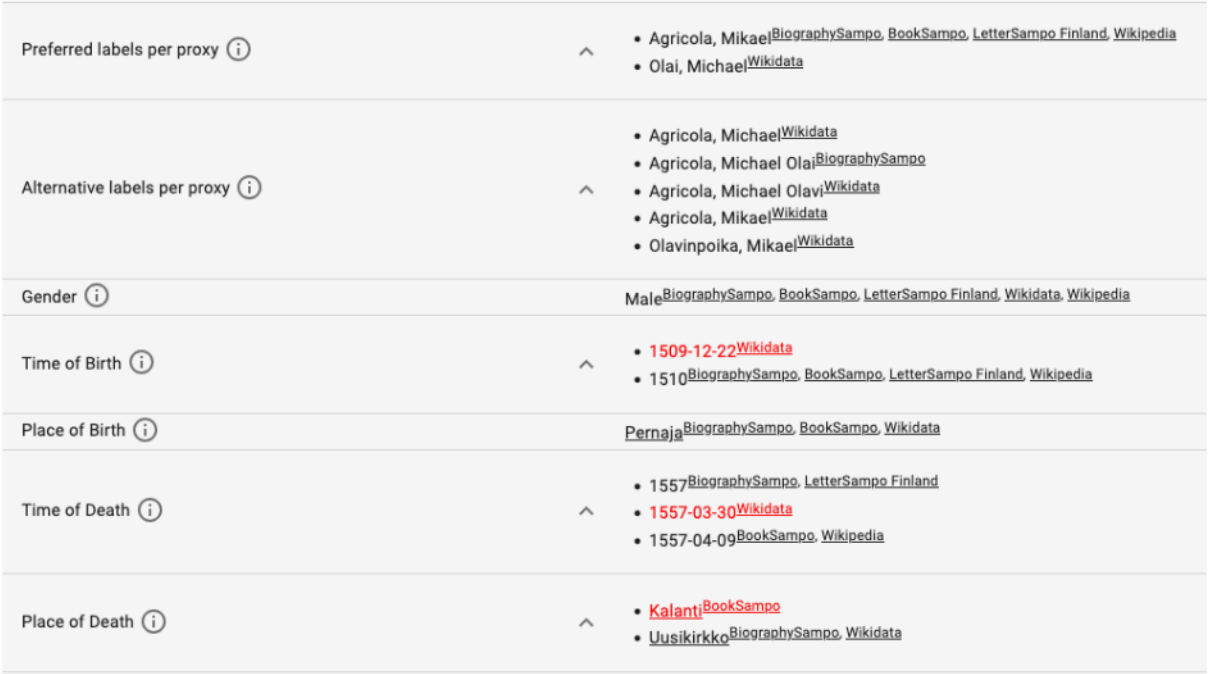

Figure 4: Screenshot of the portal showing the biographical data of Finnish bishop Mikael Agricola (1510–1557).

Table 2
Overview of inconsistencies in geographical data

	latitude			longitude		
	#Total	#Inc.	%	#Total	#Inc.	%
BiographySampo	2567	125	4.87	2567	144	5.61
AcademySampo	4481	105	2.34	4481	136	3.04
ParliamentSampo	4388	97	2.21	4388	124	2.83
BookSampo	6404	62	0.97	6404	75	1.17
Wikidata	20121	45	0.22	20121	56	0.28
LetterSampo Finland	10787	17	0.16	10787	22	0.20
WarSampo	1641	0	0.00	1641	0	0.00
All datasets	50389	451	0.90	50389	557	1.10

Canada has three different listed latitudes—one of which is marked as inconsistent—and four different longitudes—three of which are marked as inconsistent—yet all of these coordinates fall within the contemporary borders of the country (see Figure 2.2). Different data sources used for originally getting the coordinates might apply different conventions for determining the preferred coordinate point they use, e.g., by choosing to use the coordinates of a capital of a country—which might even change over time, leading to sources using this same convention to still have different coordinates depending on data creation and modification date—or by choosing a suitable center point of a land area.

3. Conclusions

This paper presented a new approach to analyze consistency of data represented in a cloud of interlinked cultural heritage KGs. The approach is based on using a data linking service KG as a basis, and was applied to the cultural heritage linked open data cloud of the Sampo system KGs and a selection of additional external datasets. However, the method is generalizable and applicable to other similar cases, such as to the international Linked Open Data Cloud[5] and the viaf.org data service for linking data from National Libraries around the world and beyond.

[5]Linked Open Data Cloud: https://lod-cloud.net/

The approach has been implemented as part of a new prototype *SampoSampo – Connecting Everything to Everything Else*, a new "metasampo" on top of other Sampo KGs and related external data sources. Our first data-analyses using SampoSampo indicate a surprisingly large number of disagreements and issues in data quality and precision regarding the biographical and geographical information in the primary data sources involved.

The SampoSampo portal[6] and the underlying LOD service[7] will be opened using the open MIT and CC BY 4.0 licenses in 2025 or early 2026.

Acknowledgments: Our work is part of the national FIN-CLARIAH research infrastructure programme, funded by the Research Council of Finland and and the European Union – NextGenerationEU instrument under grant number 346323. The last author was funded also by an Eminentia Grant of the Finnish Cultural Foundation for reflecting the research of the SeCo research group in 2001–2025. CSC – IT Center for Science has provided computational resources for our work.

References

Došilović, Filip Karlo, Mario Brčić, and Nikica Hlupić. 2018. "Explainable artificial intelligence: A survey." In *41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)s.* IEEE.

Frosterus, Matias, Jouni Tuominen, Sini Pessala, and Eero Hyvönen. 2015. "Linked Open Ontology cloud: managing a system of interlinked cross-domain light-weight ontologies." *International Journal of Metadata, Semantics and Ontologies* 10 (3): 189–201. https://doi.org/10.1504/IJMSO.2015.073879.

Hickey, Thomas B., and Jenny A. Toves. 2014. "Managing Ambiguity In VIAF." *DLib Magazine* 20 (7/8). https://doi.org/doi:10.1045/july2014-hickey.

Hyvönen, Eero. 2022. "Digital Humanities on the Semantic Web: Sampo Model and Portal Series." *Semantic Web* 14 (4): 729–744. https://doi.org/10.3233/SW-223034.

Hyvönen, Eero, Annastiina Ahola, Petri Leskinen, Heikki Rantala, and Jouni Tuominen. 2025. "How to Create a Portal for Digital Humanities Research Using a Linked Open Data Cloud of Cultural Heritage Knowledge Graphs: Case SampoSampo." In *Proceedings of the Second International Workshop of Semantic Digital Humanities (SemDH 2025), co-located with the Extended Semantic Web Conference 2025 (ESWC 2025),* vol. 4009. CEUR Workshop Proceedings, June. https://ceur-ws.org/Vol-4009/paper_11.pdf.

Hyvönen, Eero, Annastiina Ahola, Petri Leskinen, and Jouni Tuominen. 2025. "SampoSampo: A Portal for Studying Enriched Data and Semantic Connections on a Cultural Heritage Linked Open Data Cloud." In *The Semantic Web: ESWC 2025 Satellite Events, Portoroz, Slovenia, June 1 - 5, 2025, Proceedings,* 67–74. Springer-Verlag. https://doi.org/10.1007/978-3-031-99554-5_13.

Hyvönen, Eero, Petri Leskinen, Henna Poikkimäki, Heikki Rantala, Jouni Tuominen, Senka Drobac, Ossi Koho, Ilona Pikkanen, and Hanna-Leena Paloposki. 2025. "LetterSampo Finland (1809–1917) Data Service and Portal: Searching, Exploring, and Analyzing Historical Letters and Their Underlying Networks." In *Proceedings of ESWC 2025, supplement, poster and demo papers,* Accepted, forthcoming. Springer-Verlag. https://seco.cs.aalto.fi/publications/2025/hyvonen-et-al-lettersampo-finland-poster-2025.pdf.

Isaac, Antoine. 2023. *Europeana Data Model Primer.* Technical report. Europeana. https://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Primer_130714.pdf.

[6] SampoSampo portal: https://samposampo.ldf.fi/
[7] SampoSampo LOD service: https://ldf.fi/dataset/ss/

Laouenan, Morgane, Palaash Bhargava, Jean-Benoıt Eyméoud, Olivier Gergaud, Guillaume Plique, and Etienne Wasmer. 2022. "A cross-verified database of notable people, 3500BC-2018AD." *Scientific Data* 9 (1): 290. https://www.nature.com/articles/s41597-022-01369-4.pdf.

Xia, Weiguo, Ernesto Jiménez-Ruiz, and Valerie V. Cross. 2015. "Using BioPortal as a Repository for Mediating Ontologies in Ontology Alignment." In *Workshop on Semantic Web Applications and Tools for Life Sciences*. https://api.semanticscholar.org/CorpusID:37359417.